Задание для защиты итоговой аттестации

Общие требования к оформлению

На защиту необходимо подготовить презентацию в формате pdf и архив со всеми данными и скриптами.
Презентация должна содержать следующую информацию:

- Цели и Задачи работы
- Исследование (анализ литературы, существующие решения, сложности и современное состояния данного вопроса)
- Презентация решения
- Выводы
- Рекомендации по улучшению и дальнейшему развитию

Презентация должна содержать не более 15 слайдов. Само выступление должно быть не более 10 мин
Архив должен содержать:

- Файл readme с описанием всех скриптов и инструкцией по использованию проекта. Пример хорошего readme https://gist.github.com/bzvyagintsev/0c4adf4403d4261808d75f9576c814c2
- Все данные, которые были использованы в проекте. Данные не должны содержать личные характеристики реальных людей
- Скрипты, настроечные файлы

Задание

Рекомендуем исследование оформлять в jupyter notebook с текстовыми пояснениями и выводами (маркдаунами). Код без комментариев и выводов не принимается к защите.

1. Для проведения исследования требуется выбрать данные, которые не содержат персональных характеристик реальных людей. Рекомендуем выбирать данные из наборов, доступных на Kaggle или других источниках данных. Использование выбранного набора данных необходимо согласовать с вашим наставником.
2. Проведение разведочного анализа данных (EDA)
    a. Необходимо рассчитать статистики (например, медиана, дисперсия, квантили и так далее)
    b. Построить полезные графики
    c. Сделать выводы
3. Предобработка данных.
    a. Если в данных есть пропущенные значения, необходимо их обработать
    b. Если в данных есть выбросы, необходимо их обработать
    c. Если в данных есть категориальные значения, необходимо их обработать

    d. Если признаков очень много, воспользуетесь методами отбора признаков или методами понижения размерности (например, PCA)
4. Проверка статистических гипотез.
    a. Необходимо сформулировать на данных минимум 2 гипотезы и проверить их с помощью статистических критериев
5. Построение моделей машинного обучения. Необходимо построить минимум 5 моделей машинного обучения с использованием следующих алгоритмов (необходимо использовать разные алгоритмы)
    a. Линейная регрессия/логистическая регрессия
    b. Метод knn
    c. Дерево решений
    d. Random forest
    e. Градиентный бустинг (можно любой фреймворк, либо sklearn, xgboost, lightgbm, catboost)
6. Сравнить полученные модели с помощью метрик качества. Выбрать лучшую модель. Сделать выводы. Выгрузить лучшую модель из блокнота jupyter (например, с помощью библиотеки joblib или pickle).
7. Описать как планируется использовать модель после выгрузки. Создать еще один блокнот jupyter (или код на языке python в файле *.py) и загрузить модель для использования (например, с помощью библиотеки joblib). Должна быть реализована следующая логика:
    a. Пользователь вводит данные
    b. Далее запускаем ячейку в jupyter notebook или запускаем файл *.py
    c. Получаем ответ модели и рекомендации

Дополнительно. Возможна реализация использования модели не в jupyter notebook или файле python *.py, а с применением библиотеки Streamlit (или аналогичных).

Шкала оценивания:

Оценки 1/«отлично» заслуживает работа, в которой полностью и всесторонне раскрыто содержание программы обучения, обоснован выбор модели, представлен работающий код, содержится творческий подход к решению вопросов, сделаны обоснованные предложения и на все вопросы при защите слушатель дал аргументированные ответы. Проект соответствует указанным показателям.

Оценки 0.8/«хорошо» заслуживает работа, в которой содержание изложено на высоком уровне, правильно сформулированы выводы и даны обоснованные предложения, на все вопросы слушатель дал правильные ответы. Проект в большей степени соответствует указанным показателям.

Оценки 0.5/«удовлетворительно» заслуживает работа, в которой в основном раскрыто содержание программы обучения, выводы в основном правильные. Предложения представляют интерес, но недостаточно аргументированы и на все вопросы слушатель дал правильные ответы. Проект в целом соответствует указанным показателям.

Оценки 0/«неудовлетворительно» заслуживает работа, которая в основном раскрывает поставленную тему, но при защите слушатель не дал правильных ответов на большинство заданных вопросов, то есть обнаружил серьезные пробелы в профессиональных знаниях, либо в проекте не проведено ни одного эксперимента. Проект не соответствует указанным показателям.